Published online 10 November 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D109–D114***
doi:10.1093/nar/gkr988

KEGG for integration and interpretation of large-scale molecular data sets

Minoru Kanehisa[1,2,*], Susumu Goto[1], Yoko Sato[3], Miho Furumichi[1] and Mao Tanabe[1]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, [2]Human Genome Center, Institute of Medical Science, University of Tokyo, Minato-ku, Tokyo 108-8639 and [3]Life Science Solutions Department, Fujitsu Kyushu Systems Ltd., Sawara-ku, Fukuoka 814-8589, Japan

Received September 15, 2011; Accepted October 17, 2011

ABSTRACT

Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/ or http://www.kegg.jp/) is a database resource that integrates genomic, chemical and systemic functional information. In particular, gene catalogs from completely sequenced genomes are linked to higher-level systemic functions of the cell, the organism and the ecosystem. Major efforts have been undertaken to manually create a knowledge base for such systemic functions by capturing and organizing experimental knowledge in computable forms; namely, in the forms of KEGG pathway maps, BRITE functional hierarchies and KEGG modules. Continuous efforts have also been made to develop and improve the cross-species annotation procedure for linking genomes to the molecular networks through the KEGG Orthology system. Here we report KEGG Mapper, a collection of tools for KEGG PATHWAY, BRITE and MODULE mapping, enabling integration and interpretation of large-scale data sets. We also report a variant of the KEGG mapping procedure to extend the knowledge base, where different types of data and knowledge, such as disease genes and drug targets, are integrated as part of the KEGG molecular networks. Finally, we describe recent enhancements to the KEGG content, especially the incorporation of disease and drug information used in practice and in society, to support translational bioinformatics.

INTRODUCTION

The large-scale molecular data sets generated by genome sequencing and other high-throughput experimental technologies are the basis for understanding life as a molecular system and for developing practical applications in medical, pharmaceutical and environmental sciences. The key to linking such large-scale data sets to practical values lies in bioinformatics technologies, not only in terms of computational methods, but also in terms of knowledge bases. Since 1995, we have been developing Kyoto Encyclopedia of Genes and Genomes (KEGG), a reference knowledge base for deciphering the genome. Experimental knowledge on systemic functions of the cell and the organism is represented in terms of molecular networks (KEGG pathway maps, BRITE functional hierarchies and KEGG modules), and a mechanism (KEGG Orthology system) is developed for linking genes in the genome to nodes of the molecular network. Over the past 16 years, KEGG has been expanded significantly to meet the needs of both large projects and individual laboratories. In recent years, our efforts have been focused on capturing and representing knowledge on diseases as perturbed states of the molecular network and drugs as perturbants to the molecular network (1).

The scientists' promises regarding the medical and social benefits of the Human Genome Project and subsequent projects appear to be long overdue, but thanks to the advancement of next-generation sequencing technologies the time may have finally come for bringing the genomic revolution to society. To meet the needs of translational bioinformatics the KEGG resource is being integrated with disease and drug information used in clinical practice and in wider society, such as the package inserts of marketed drugs. It is hoped that KEGG will assist scientists to translate their research results into medical and industrial innovations and will also allow doctors, pharmacists, patients and consumers to make better use of scientific data on disease and drug-related molecular networks.

OVERVIEW OF KEGG

KEGG databases

KEGG is an integrated database resource consisting of the 15 main databases shown in Table 1. It covers various data objects, called KEGG objects, for computer

*To whom correspondence should be addressed. Tel: +81 774 38 3270; Fax: +81 774 38 3269; Email: kanehisa@kuicr.kyoto-u.ac.jp

representation of molecular systems, broadly categorized into systems information (PATHWAY, BRITE, MODULE, DISEASE, DRUG and ENVIRON), genomic information (ORTHOLOGY, GENOME and GENES) and chemical information (COMPOUND, GLYCAN, REACTION, RPAIR, RCLASS and ENZYME). The identifier of each database entry is generally in the form of 'db:entry' where 'db' is the database name and 'entry' is the entry name. However, 'db' may be omitted in 13 of the listed databases, because the entry name, called the KEGG object identifier consisting of a database-dependent prefix and a five-digit number, is unique across the databases. These 13 databases are all manually created by KEGG. The remaining two databases, KEGG GENES derived from RefSeq (2) and KEGG ENZYME derived from ExplorEnz (3), are also given KEGG-original annotations. In the genomic information category there are auxiliary databases that are computationally generated, including KEGG DGENES for genes in draft genomes, KEGG EGENES for genes as EST contigs (4), newly introduced KEGG MGENES for genes in metagenomes, and KEGG SSDB for sequence similarity relationships among KEGG GENES (5).

KEGG organisms

KEGG GENES is a collection of gene catalogs for all organisms whose genomes are completely sequenced and made available by RefSeq. Each organism is treated like a database with the three-letter organism code as the database name and the gene identifier (either NCBI Gene ID or locus_tag) as the entry name. The three-letter organism code is an alias of the T number identifier in the KEGG GENOME database; for example, 'hsa' for *Homo sapiens* is equivalent to 'T01001' (Table 1). The organism code is also used as a prefix to identify organism-specific versions of KEGG pathway maps, BRITE hierarchies and KEGG modules (see 'KEGG Orthology' section). The organisms of KEGG DGENES and EGENES may be identified with the four-letter organism codes starting with 'd' and 'e', respectively, or the T numbers. The environmental samples of KEGG MGENES are identified only by the T numbers.

KEGG pathway maps

The KEGG pathway maps are graphical diagrams representing knowledge on molecular interaction and reaction networks for metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases and drug development. Each map is manually drawn with in-house software called KegSketch, summarizing experimental evidence in published literature. Basic graphics objects in the KEGG pathway maps are: (i) boxes for KEGG Orthology (KO) groups identified by K numbers, (ii) circles for small molecules identified by C numbers, glycans identified by G numbers, and other molecules, and (iii) lines for KO groups in global metabolism maps. Boxes in regular metabolic maps and lines in global metabolic maps, both for KO groups, are also linked to enzymes identified by EC numbers and reactions identified by R numbers. The reference pathway map identified with prefix 'map' contains all three, and the separate version of the reference pathway map is identified with prefix 'ko', 'ec', or 'rn' (Tables 1 and 2).

BRITE functional hierarchies

The BRITE hierarchy files, called the htext (hierarchical text) files, represent known functional hierarchies (hierarchical classifications) of genes and proteins, diseases and drugs, compounds and reactions, and organisms and cells. Each htext file is manually created with in-house software called KegHierEditor. It contains 'A', 'B', 'C', etc. at the first column to indicate the hierarchy level and may contain multiple tab-delimited columns. Thus, the htext file is like an Excel file with an additional first column for the hierarchy level. There are two classes of BRITE htext files (Tables 1 and 2). Those identified with prefix 'ko' are for gene and protein classifications represented by KO groups (K numbers), and those identified with prefix 'br' (or Japanese version 'jp') are

Table 1. KEGG databases and KEGG object identifiers

Category	Database	Content	Prefix	Example
Systems information	KEGG PATHWAY	Pathway maps	map, ko, ec, rn, (org)	hsa04930
	KEGG BRITE	Functional hierarchies	br, jp, ko, (org)	ko01003
	KEGG MODULE	KEGG modules	M, (org)_M	M00008
	KEGG DISEASE	Human diseases	H	H00004
	KEGG DRUG	Drugs	D	D01441
	KEGG ENVIRON	Crude drugs, etc.	E	E00048
Genomic information	KEGG ORTHOLOGY	KO groups	K	K04527
	KEGG GENOME	KEGG organisms	T	T01001 (hsa)
	KEGG GENES	Genes in high-quality genomes		hsa:3634
Chemical information	KEGG COMPOUND	Metabolites and other small molecules	C	C00031
	KEGG GLYCAN	Glycans	G	G00109
	KEGG REACTION	Biochemical reactions	R	R00259
	KEGG RPAIR	Reactant pairs	RP	RP04458
	KEGG RCLASS	Reaction class	RC	RC00046
	KEGG ENZYME	Enzyme nomenclature		ec:2.7.10.1

'(org)' represents three-, four- or five-letter organism code.

Table 2. Naming convention of KEGG molecular networks

Molecular network	Manually created		Computationally generated		
	Reference		Organism-specific	Disease/drug	Other
Metabolic pathway	map00010	ko00010 ec00010 rn00010	hsa00010	hsadd00010	
Non-metabolic pathway	map02010	ko02010	hsa02010	hsadd02010	
Gene/protein brite	ko02000		hsa02000	hsa02000_dd	
Non-gene/protein brite	br08303				br08303_target br08303_enzyme
Module	M00001		hsa_M00001		

Five-digit numbers are examples only.

for other classifications, such as diseases (H numbers), drugs (D numbers) and compounds (C numbers).

KEGG modules

KEGG MODULE was originally introduced to define tighter functional units than KEGG PATHWAY (6) so that the pathway information in KEGG would be represented in three resolutions: global maps (for metabolism), regular maps and modules. Subsequently, the scope of KEGG MODULE has been extended and there are currently four types of KEGG modules: (i) pathway modules, (ii) structural complexes, (iii) functional sets and (iv) signature modules. The first three types of modules usually correspond to parts of KEGG pathway maps and BRITE hierarchies. The signature module is a set of genes in the genome, or perhaps in the transcriptome as well, that can be used as a marker for the phenotype, such as pathogenicity and metabolic capacity. Each KEGG module is defined by the combination of K numbers and associated with an automatically generated module map according to the predefined notations: space delimited items for pathway elements, comma separated items in parentheses for alternatives, plus sign to define a complex and minus sign for an optional item.

KEGG Orthology

The KEGG pathway maps, BRITE functional hierarchies and KEGG modules are created in a general way to be applicable to all organisms; namely, in terms of the KO groups rather than individual gene names in specific organisms. Each KO entry represents a manually defined and context-dependent ortholog group that generally corresponds to a node of the pathway map, BRITE hierarchy or KEGG module, and which consists of orthologous genes in all available genomes assigned by the genome annotation procedure described below. Thus, the organism-specific versions of pathway maps, BRITE hierarchies and KEGG modules can be automatically generated by converting K number nodes to gene identifier nodes as well as by coloring in green (which is a KEGG convention). The organism-specific version is identified by using the organism code in place of 'ko' as a prefix in PATHWAY and BRITE, and in addition to the M number prefix in MODULE as shown in Table 2.

Genome annotation in KEGG

The genome annotation in KEGG is essentially cross-species annotation, finding orthologous genes in all available genomes for given K numbers, and is currently performed as follows. (i) Experimental evidence on known functions of genes and proteins is organized in the KO database, which is created together with the KEGG PATHWAY, BRITE and MODULE databases. (ii) Gene catalogs of complete genomes are generated from RefSeq and other public resources. (iii) All pairs of genomes (gene catalogs) are compared by the SSEARCH program, and the GFIT tables are generated detailing the information for each gene in a genome about best-hit genes in all other genomes (5). (iv) GFIT tables are continuously updated, and the automatic version of the KOALA tool (1) presents to human annotators a summary of discrepancies between its K number assignment and the current annotation. (v) Discrepancies are examined by annotators with the manual version of KOALA and other tools such as for protein domains, ortholog tables and gene clusters. We plan to incorporate KEGG modules into the KOALA annotation procedure.

NEW FEATURES OF KEGG

KEGG Mapper

KEGG PATHWAY, KEGG BRITE and KEGG MODULE constitute the KEGG reference knowledge base for biological interpretation of molecular data sets, especially large-scale data sets generated by high-throughput experimental technologies. This is accomplished by the process of KEGG mapping, which is to map, for example, a genomic content of genes to generate organism-specific versions of pathways as mentioned above or a transcriptomic content of genes to see which parts of pathways are up/down regulated. The mapping process is considered a set operation between the query data set and the target data of the KEGG reference knowledge base.

KEGG Mapper is the user interface for KEGG mapping. It currently consists of seven tools as shown in Table 3. For the three basic tools (Search Pathway, Search Brite and Search Module) the query data set is a collection of molecular objects (genes, proteins, small molecules,

etc.), and the user would simply query the presence or absence of objects, resulting in pathway and ontology enrichment. In the advanced tools (Search&Color Pathway, Search&Color Brite, Color Pathway and Join Brite) the query data can consist of molecular objects and their attributes; for example, genes with up/down expression levels and genes with disease names known to be associated. The attributes need to be processed and color-coded when using the Search&Color Pathway and Search&Color Brite tools. The Color Pathway tool accepts numerical values, which may be displayed by color shading or by overlaying 3D bar graphs on top of the KEGG pathway map. Figure 1 shows an example of the latter type of display, where genes in a cancer pathway are associated with somatic mutation frequency. The Join Brite tool accepts any text information and returns the result by adding a column or another hierarchy level in the BRITE hierarchy file (see: Knowledge base extension).

Genome comparison and combination

For the organisms (GENES, DGENES and EGENES) and environmental samples (MGENES) available in KEGG, it is now possible to map multiple data sets against KEGG pathway maps and BRITE functional hierarchies. The user interface may be found in the KEGG GENOME page, and the result is displayed using multiple color coding. This feature can be used, for example, to compare metabolic capabilities of different organisms (see Figure 2 for human and *Escherichia coli* comparison), to examine complementarity of host–symbiont, host–pathogen and host–microbiome relationships, and to examine collective features of pangenomes. These tasks may be applied to the user's own genomes by using KEGG Mapper with preprocessing of K number assignment and color specification.

Table 3. KEGG Mapper tools

Tool	Query data	Reference knowledge
Search Pathway	Objects	KEGG PATHWAY database
Search&Color Pathway	Object–attributes relations	KEGG PATHWAY database
Color Pathway	Object–attributes relations	KEGG PATHWAY map
Search Brite	Objects	KEGG BRITE database
Search&Color Brite	Object–attributes relations	KEGG BRITE database
Join Brite	Object–attributes relations	BRITE functional hierarchy
Search Module	Objects	KEGG MODULE database

Knowledge base extension

The query data in KEGG Mapper is usually raw and uninterpreted data, but the KEGG mapping set operation may also be applied to well curated data sets. In fact it was the original concept of the KEGG project to automatically generate organism-specific pathways through the set operation between manually annotated genome data and manually created pathway maps. We have extended this approach as more data on diseases and drugs are accumulated in KEGG, effectively enhancing the KEGG knowledge base and enabling KEGG Mapper and other applications to easily integrate disease and drug information.

KEGG DRUG is a comprehensive collection of approved drugs in Japan, USA and Europe unified based on chemical structures and/or chemical components. It contains rich information about molecular networks (such as targets, metabolizing enzymes and drug–drug interactions) enabling integrated analysis with KEGG pathway maps and BRITE hierarchies. KEGG DISEASE is still an underdeveloped database, but it aims to computerize all human diseases with known genetic factors and all infectious diseases with known pathogen genomes, so that diseases can be analyzed as perturbed states of the molecular networks.

The mapping of disease and drug data from these databases is now incorporated in the daily KEGG update procedure. First, all known disease genes accumulated in KEGG DISEASE and all known drug targets accumulated in KEGG DRUG are integrated in the

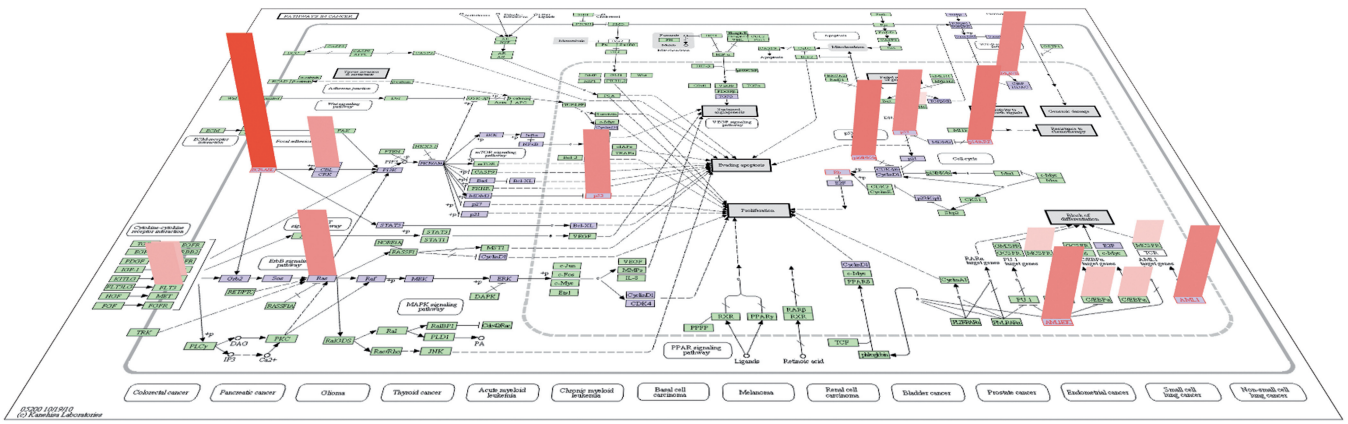

Figure 1. The Color Pathway tool in KEGG Mapper accepts numerical values associated with genes, which can be displayed by color-shading or 3D mapping. Here somatic mutation frequency observed in chronic myeloid leukemia (obtained from the COSMIC database, http://www.sanger.ac.uk/genetics/CGP/cosmic/) is shown on top of the global cancer pathway map (hsa05200).

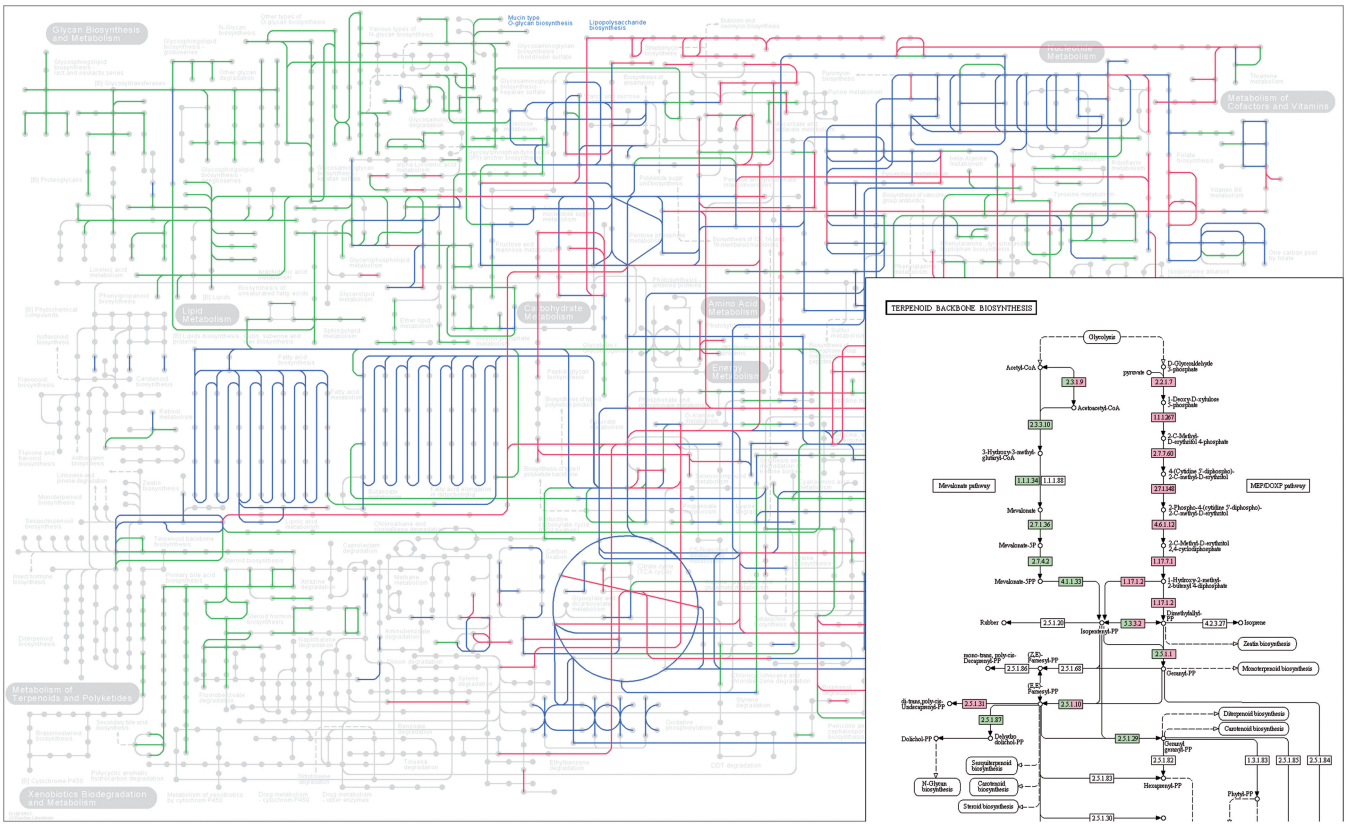

Figure 2. Comparison of metabolic pathways reconstructed from the complete genomes of *Homo sapiens* (hsa) and *E. coli* (eco). Here the global metabolic pathway map (map01100) and the terpenoid backbone biosynthesis map (map00900) are shown, where human-specific pathways are colored in green, *E. coli* specific pathways in pink, and shared pathways in blue or split coloring.

KEGG PATHWAY and BRITE databases. This is accomplished by preparing binary relations, human gene identifier to H number and human gene identifier to D number, as query data sets and applying the Search&Color Pathway and Search&Color Brite tools. The generated pathway maps and BRITE hierarchy files are identified by the 'hsadd' prefix and the '_dd' modifier, respectively, as shown in Table 2. The mapping result is displayed by coloring: pink when the gene is associated with a disease and light blue when the gene product is a drug target. Figure 3 shows this disease/drug mapping against the Alzheimer's disease pathway map.

Second, drug targets, drug metabolizing enzymes and other drug interaction data accumulated in KEGG DRUG are mapped to selected BRITE hierarchy files by the Join Brite tool. For example, drug targets and drug metabolizing enzymes are mapped to the WHO's ATC (Anatomical Therapeutic Chemical) classification, and the generated versions, 'br08303_target' and 'br08303_enzyme' (Table 2), contain additional information in an additional column. The mapping result would allow complementation of missing data and prediction of new data, as similar chemical substances within this functional hierarchy would have similar properties.

Translational bioinformatics

The KEGG mapping approach is being further expanded to integrate disease and drug information used in practice and in society, as well as related data contained in outside resources. The Japanese version of the KEGG DRUG database is already tightly integrated with the package insert information of all prescription drugs and OTC drugs marketed in Japan, provided by Japan Pharmaceutical Information Center (JAPIC) and served by GenomeNet (http://www.genome.jp/kusuri/). This is reflected, for example, in the Japanese version of the ATC classification file, 'jp08303_japic', which contains product names in an additional hierarchy level. A similar integration will be performed for the drug products marketed in USA and other countries.

Using the package inserts of all prescription drugs in Japan, we extracted adverse drug–drug interactions from the description of contraindications and warnings, converted them into KEGG DRUG D number pairs and tried to categorize interaction patterns, such as those involving metabolizing enzymes or overlapping targets (7). The DDI (drug–drug interaction) search based on this data set is available in each KEGG DRUG entry page, and the result may be displayed on top of the ATC hierarchy.

While the main focus of the KEGG resource has been the molecular network, the next step will have to include the association of network states and phenotypic features, such as network perturbations associated with diseases, drug effects and other health conditions. The use of controlled vocabulary to represent phenotypic features,

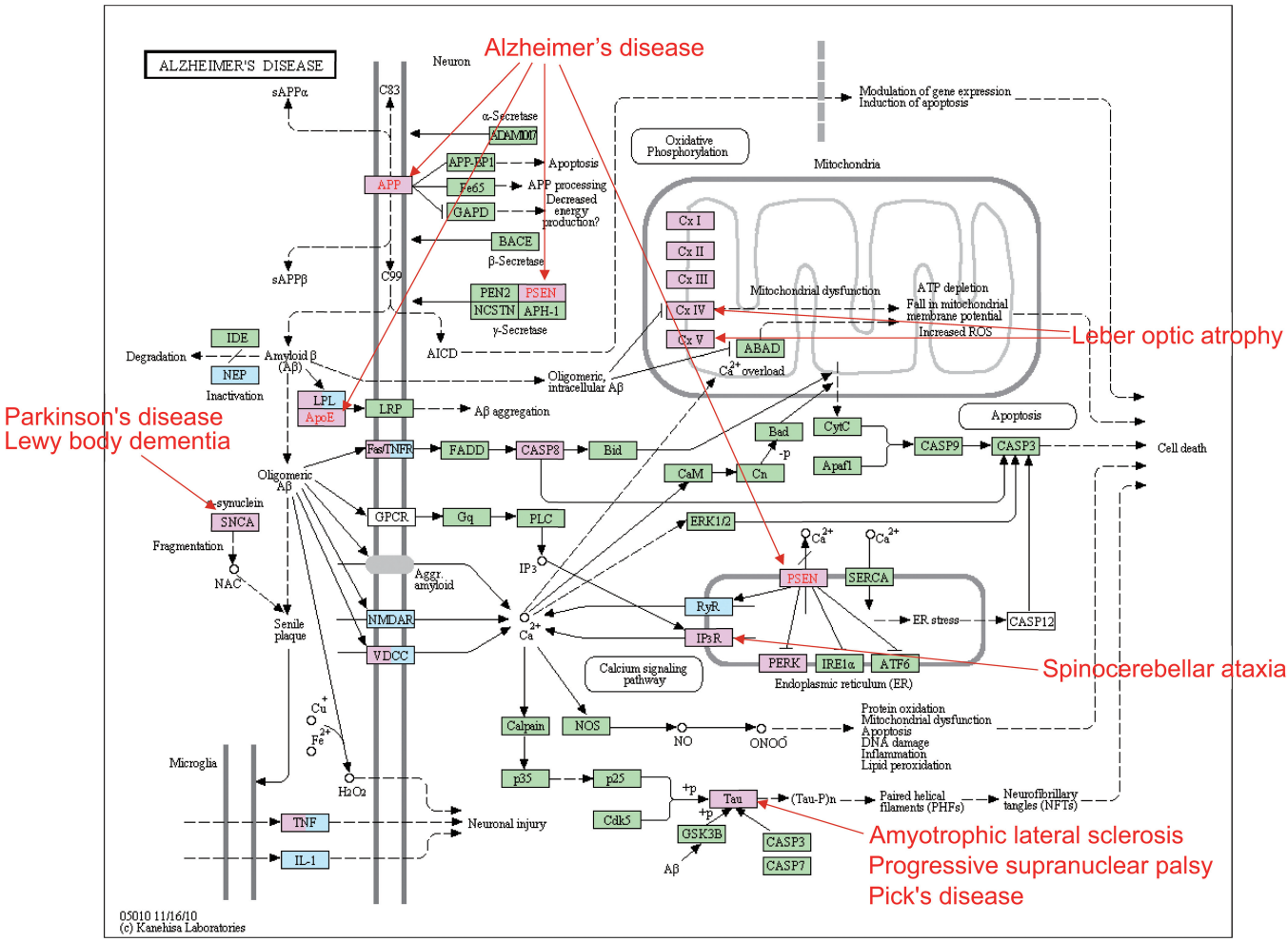

Figure 3. Disease/drug mapping is the process to map all known disease genes (pink) and all known drug targets (light blue) against all KEGG pathway maps. This is the mapping result of the Alzheimer's disease pathway map (hsa05010), which reveals relationships with other neurodegenerative diseases.

like that found in MedDRA (http://www.meddramsso.com/), is an ongoing project.

Accessing KEGG

KEGG is made available at both the GenomeNet website (http://www.genome.jp/kegg/) and the KEGG website (http://www.kegg.jp/). The GenomeNet site has been the primary site, but this will change at the end of 2011. The KEGG site will then be the primary site handling all database update procedures, and the GenomeNet site will become a mirror site.

FUNDING

Japan Science and Technology Agency (KEGG project, partial); Bioinformatics Center, Institute for Chemical Research, Kyoto University (computational resource). Funding for open access charge: Kyoto University Bioinformatics Center.

Conflict of interest statement. None declared.

REFERENCES

1. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.
2. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res*, **37**, D32–D36.
3. McDonald,A.G., Boyce,S. and Tipton,K.F. (2009) ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res.*, **37**, D593–D597.
4. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
5. Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.
6. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
7. Takarabe,M., Shigemizu,D., Kotera,M., Goto,S. and Kanehisa,M. (2011) Network-based analysis and characterization of adverse drug-drug interactions. *J. Chem. Inf. Model*, October 11 (doi:10.1021/ci200367w; epub ahead of print).